A comparative framework for understanding clustering results

Christopher Hart, Benjamin Bornstein, Joseph Roden, Barbara Wold, Eric Mjolsness*
Division of Biology and/or Jet Propulsion Laboratory
California Institute of Technology

* = presenting

Unsupervised classification techniques are frequently employed to partition and organize large scale gene expression results. We have constructed both a mathematical framework and a software infrastructure to quantitatively compare the results of such machine learning techniques. In order to gain a better understanding of the behaviors of a variety of commonly used algorithms as a function of data structure and algorithm parameters we evaluated several algorithm's ability to recapitulate the structure of various synthetic data sets using this comparative framework. Our experiments using this framework provides suggestions of which algorithms should perform best on data of a given structure. This framework also provides a toolset which can be useful in combining the results of several clusterings.

Initial results suggest that K-means is particularly sensitive to the algorithms initialization and choice of cluster number. K-means also appears to perform poorly as data becomes more complex. Self Organizing Maps (SOMs) and a novel agglomerated version of a phylogenic tree perform much more consistently across varying data structures and parameter sets. Most surprising was the ability of each of these algorithms to perform well even when the choice of cluster number was much greater than the true number of clusters.